Docket No. D3239-00002

METHOD, SYSTEM AND COMPUTER PROGRAM FOR REDACTION OF MATERIAL FROM DOCUMENTS

Field of the Invention.

This invention relates to computer programs for redaction of material from documents stored as electronic files.

Background of the Invention.

For a variety of reasons, documents must be presented in their original format, with certain material, such as text or illustrations, redacted. For example, in response to Freedom of Information Act requests, U.S. Government agencies must release certain documents. Often, these documents contain information which is exempt from disclosure under FOIA. The documents themselves are released in their original format, but the material that is exempt from disclosure must be removed. In other situations, such as in litigation, counsel may wish to provide a document in response to a discovery request, but remove privileged portions of the document.

Traditionally, the primary method of redacting information from documents was to take a physical copy of the document, and cover over the text and images to be redacted. Covering would be done manually with tools such as a black marker, a grease pencil, or strips of opaque tape. This process is slow and labor intensive. Also, on review of the redacted document, if the reviewer determines that information which is required to be disclosed has been redacted, it is difficult to correct such errors. Once the process is completed, the document can be photocopied and a photocopy furnished to the requester. Many U.S. Government agencies have been placing

such documents on World Wide Web servers to make the documents available over the Internet. The documents must be scanned before being placed on the Web server.

Methods have been developed to perform redaction of electronic documents. Typically, these methods operate on tag image file format (TIFF) files. Conventionally, these products employ overlays of opaque blocks of color to cover information in the TIFF document.

A variety of disadvantages are associated with TIFF files, particularly when compared with the portable document format (PDF) files in the format developed by Adobe Systems, Inc. For example, in the TIFF format, designers have great flexibility in creating tags, which results in certain applications not being able to read all TIFF files. More recent versions of TIFF are not interchangeable with older versions. Also TIFF files for the same document are much larger than comparable PDF files. The overlays of opaque blocks of color, ordinarily saved separately from the document pages, causes a delay when a document is opened while the overlays are placed by software on the page. Generally, custom programming and system integration is required for the TIFF programs to be operational, because not all applications are compatible with TIFF files. If the overlays of blocks of color are not properly embedded in the document, it is possible for the overlays to be removed. The software packages that are commercially available for redaction of TIFF files are only available for certain operating systems, and do not operate on a variety of platform. For example, such systems are not compatible with both systems that operate on Windows PC's and Macintosh PC's. Finally, if large portions of the text are removed, the opaque blocks of color tend to be very large. This can render printing of the documents to be problematic.

Objects and Advantages of the Invention.

It is an object of the invention to provide a method, a system, and a computer program for convenient identification of portions of a document to be redacted, and for rapid redaction of electronic documents.

It is a further object of the invention to provide a method, system, and computer program for redaction of electronic documents that eliminates the risk of unauthorized recovery from the redacted document of the material that was excised.

Among the advantages of the invention are that the foregoing objects are achieved.

Additional objects and advantages of the invention will become evident from the detailed description of a preferred embodiment, which follows.

Summary of the Invention.

A method according to the invention includes the steps of selecting a geometric area on the document for redaction and representing the geometric area as one or more objects. The information in the document representing material to be reviewed, such as text and images, is then represented as a series of geometric locations and codes and stored as one or more objects. The objects representing the geometric area and the objects representing the document information are then compared. An output is created in which the document information is replaced where the geometric location of the geometric areas corresponds to the geometric location of the document information. The resulting output does not contain any of the removed information.

Brief Description of the Figures.

Figure 1 depicts a computer system according to the invention.

Figure 2 is a sample screen illustrating a step in a method according to the invention.

Figure 3 is a flow chart illustrating steps and a method according to the invention.

Figure 4 is also a sample screen illustrating a step in a method according to the invention.

Figure 5 is a sample screen illustrating a document redacted according to the invention.

Detailed Description of An Embodiment of the Invention.

Referring to Figure 1, there is depicted schematically the principal hardware components of a computer system 10 of a type which may be used in a method and system according to the invention. Computer system 10 has a central processing unit 15, a display 20, inputs 25, and a storage device 30. Central processing unit 15 may be any one of numerous conventional microprocessors, such as the Intel Pentium processor or various Motorola processors employed in Macintosh personal computers. Display 20 may be a CRT, LCD, or other display device. Inputs 25 are conventionally a keyboard and a mouse, although other input devices may be employed. Storage device 30 may include any storage medium on which a computer program may be stored, including without limitation hard drive, floppy drives, tapes, and the like. Of course, storage device 30 may be accessed over a communications link or network.

Referring to Figure 2, there is shown a display 20 during a step while a program according to the invention stored on a storage device is running on CPU 15. A program for the display and manipulation of portable document format (PDF) files is loaded from memory and begins running on the CPU. The PDF display program may be Adobe Acrobat, for example. The user opens the PDF file containing the document to be redacted. A redaction program

according to the invention is then loaded and begins running. The user has an option of drawing a geometric shape to start the selection process. As may be seen in Figure 2, document 40 is displayed on display 20. The redaction program displays a rectangle 45 superimposed over a portion of document 40. The user may manipulate the size and location of rectangle 45 by suitable features in the software, such as movement of a cursor 50 using a mouse or other input device. The rectangle may also be referred to as a box or a frame. In principle, other geometric shapes may be employed.

The user may select a particular size and location for rectangle 45 at any time. When the user does this, the redaction program stores an object in memory having geographic information representing the location of the borders of rectangle 45 on document 40. Referring to Figure 3, this step is indicated by block 300. The format of the geographic information is a series of coordinate points relative to the PDF page drawing coordinate system. The objects may also contain information selected by a user employing a menu or other options embodied in the redaction program. Such information may include codes identifying the nature of the redaction, or comments associated with the redaction. An example of the display of such codes is shown in Figure 4. In Figure 4, document 40 is shown with several boxes 45 representing geographic areas to be redacted. As can be seen, codes 50, such as (j)(2), (k)(1), and the like, are shown associated with boxes 45. For the convenience of the user, the content, both text and images, within the boxes, is also visible. The objects are stored in association with the document file. When another user next opens the document file, the annotations may be viewed in combination

with the file. In this manner, a first line reviewer may create annotations for review and ready change by a manager.

After any further changes have been made, the next step is the performance of the redaction. A preliminary step is the translation from the geometric space occupied by the rectangles, as reflected in the stored annotations, to the layout of the textual content of the PDF format document. The first step in the translation is the creation of a model of the content, as shown by the box labeled 305. The model is created by parsing the PDF stream. In the parsing process, the instructions, or information that relate to the text and images and the geographic location of the text and images are identified and stored. For text, these instructions include those that relate to font selection, font size, the transformation matrix through which individual characters are drawn, character and word spacing, escaped characters, and the tokens that cause text to be emitted. The resulting model represents the text as a list of objects describing each occurrence of text using font characteristics and geometric parameters. The model is stored in a convenient memory location. The objects will be referred to as text occurrence objects. Similarly, image information is modeled as a list of objects describing each image. The objects with image information may be referred to as image occurrence objects. Text occurrence objects and image occurrence objects may also be referred to more generally as content objects. Information regarding text and images in a particular geographic location on the displayed document may be dispersed in a PDF stream. The creation of the model places the information in proximity in a document.

The content objects are then compared to the annotation objects. For text occurrence objects, in the comparison process, geometric intersections identified, as indicated by the box labeled 310. It will be understood that text associated with geometric intersections is to be redacted. The program can create an output file. For text that is not identified as corresponding to an annotation object, the text object is reproduced in the output file. As text to be redacted is identified, the text is replaced in an output file with appropriate material, as indicated by the block labeled 315. The replacing material may be exemption codes, hyphens, or other material selected by the user. The output file is also in the form of objects. The output file comprises a geometric representation as a series of objects of the textual content, with those portions to be redacted removed. The redacted text is never contained in the output file that constitutes the newly-created document.

If an image is to be redacted, the program carries out the redaction on a pixel-by-pixel basis. All pixels that have been identified as to be redacted are replaced. For example, all pixels to be redacted may be changed to solid black. The redacted image information is therefore never contained in the output file where the newly-created document is found.

As an option, the program will insert additional characters in the output file for text to assure that the original width is maintained.

Figure 5 depicts a displayed redacted document 20. Selected portions of an image have been replaced by all black pixels at 55. Selected portions of the text have been replaced by hyphens, as shown at 60.

The resulting document no longer contains any of the redacted information, as the redacted information was removed in the creation of the document.

Batch annotation and redaction of documents is useful for certain users. For example, Federal government agencies may have lists of keywords that are ordinarily redacted. The method of the invention may include batch annotation and redaction. The user is provided with the option of identifying words that are to be redacted. The method then reviews a selected document in portable document format and creates annotations corresponding to each identified occurrence of the identified word. In a preferred embodiment, the PDF file is parsed and converted into text occurrence objects as described above. The program checks the text occurrence objects against the list of identified words. An annotation object is created coinciding with the geometric position of the word as shown in the text occurrence object. The annotation object can be stored in association with the PDF file. Each occurrence of each of the selected words is thereby designated for redaction. The annotation object can then be considered a candidate annotation for review by a human reviewer.

Through substantially standard techniques, a variety of information can be associated with each annotation. For example, the name of the reviewer, the time and date of creation, codes representing a reason for the exemption, and comments may be included. Displays may include different colors for different reviewers.

The program, method and system of the invention have been described above with respect to the redaction of text from a document. However, the program, method and system may be applied to content of other types, such as image information.

The computer programs in accordance with the invention may be stored and distributed in any suitable storage medium, such as fixed disk, portable diskettes, and CD-ROM or other read-only memories. Also, methods described as being carried out in software running on general-purpose computer hardware may be implemented in hardware.

It will be understood that various changes in the details, materials and arrangements of the programs, methods and systems which have been described and illustrated above in order to explain the nature of this invention may be made by those skilled in the art without departing from the principle and scope of the invention as recited in the following claims.